Smart Intrusion Detection System Comprised of Machine Learning and Deep Learning

Shah Md. Istiaque, Asif Iqbal Khan, Sajjad Waheed

Abstract — **In the present world, digital intruders can exploit the vulnerabilities of a network and are capable to collapse even a country. Attack in Estonia by digital intruders, attack in Iran's nuclear plant and intrusion of spyware in smart phone depicts the efficiency of attackers. Furthermore, centralized firewall system is not enough for ensuring a secured network. Hence, in the age of big data, where availability of data is huge and computation capability of PC is also high, there machine learning and network security have become two inseparable issues.**

In this thesis, KDD Cup'99 intrusion detection dataset is used. Total 3, 11,030 numbers of records with 41 features are available in the dataset. For finding the anomalies of the network four machine learning methods are used like Classification and Regression Tree (CART), Random Forest, Naive Bayes and Multi-Layer Perception. Initially all 41 features are used to find out the accuracy. Among all the methods, Random Forest provides 98.547% accuracy in intrusion detection which is maximum, and CART shows maximum accuracy (99.086%) to find normal flow of data. Gradually selective 15 features were taken to test the accuracy and it was found that Random Forest is still efficient (accuracy 98.266%) in detecting the fault of the network. In both cases MLP found to be a stable method where accuracy regarding benign data and intrusion are always close to 95% (93.387%, 94.312% and 95.0075, 93.652% respectively).

Finally, an IDS model is proposed where Random Forest of ML method and MLP of DL method is incorporated, to handle the intrusion in a most efficient manner.

Index Terms — **Intrusion Detection System, CART, Random Forest, Naive Bayes, Back-Propagation based MLP.**

I. INTRODUCTION

The Internet has become an important media of regular correspondence through online media collaboration, email, e-learning, and so forth. Additionally, little and large organizations have expanded their purchaser base by giving direct client showcasing, web shopping and inter organization correspondence utilizing essential web correspondence. With the gigantic development of computer network, the total system experiences security weaknesses which are troublesome and exorbitant to be solved by manufactures [1]. A few dangers are brought through utilization of incapable and wasteful security instruments welcoming intrusions from internet hackers [2]. In this way, it is clear that the prevention technologies set up like malware evacuation programs, antivirus projects and firewalls, neglect to give outright security since aggressors utilize fresher methods for ambushing the system just as its clients [3].

Over the years, operating system security technology has been upgraded to forestall the issues of confidentiality, integrity, and availability of a network [4]. Intrusion Detection System (IDS) is a prudent early warning system for a network. For the most part, IDS alarms the client before the system gear is imperiled when it identifies inner and outside intrusion [5]. At first, network manager executes IDS physically by observing the system through a console [6]. However, the objective of this research is to propose a system security framework actualizing the Artificial Neural Network (ANN), utilizing Back Propagation (BP) calculation. The examination further breaks down oddity discovery, in light of a few AI methods, using different training and testing datasets [7].

The remaining part of the paper is developed in the following way. Segment 2 provides a literature review regarding recent updates, Segment 3 highlights some classification algorithm utilized in the work, which is CART, Random Forest, Naïve Bayes and MLP. Experimental results and analysis, the core contribution of this paper, is discussed in Segment 4. Lastly, Segment 5 conclude the paper.

II. RELATED WORK

A lot of studies have already been carried out about various machine learning and deep learning methods and its application in different fields. Same is not yet exhaustively done in the field of information security. The gap found in the field of IDS was studied with available resources till today. The unique idea of applying the machine learning and deep learning methods in the IDS is a new theme in the contemporary research arena.

A. Study Regarding IDS

According to each of the detection approaches, IDS frameworks are separated into two classifications, which are anomaly-based detection and misuse based detection [8], [9]. Misuse-based IDS can recognize known assaults effectively yet neglects to discover new assaults which fail to embody the rules in the database [10]. In this manner, a database must be persistently refreshed to store the signature of each assault that is known. This IDS type is clearly incapable to identify new attacks except it is trained [11]. Anomaly based IDS can

Published on October 8, 2020.
Shah Md. Istiaque, Bangladesh University of Professionals, Bangladesh. (e-mail: sunny6358@gmail.com)
Asif Iqbal Khan, Mawlana Bhashani Science and Technology University, Bangladesh.
Sajjad Waheed, Professor Dr., Mawlana Bhashani Science and Technology University, Bangladesh.

assemble a typical conduct model and recognizes any significant deviations from the model similar to an interruption. This IDS type can identify new assaults or obscure one. However, it includes a high pace of false alarms [12], [13].

B. Study Regarding Dataset

The most significant challenge in assault identification framework is whether to produce genuine system traffic or to utilize the accessible benchmark datasets. There is criticism about the use of datasets acquired from genuine system traffic as it makes greater uncertainty and there is no such methodology that obviously discloses how to precisely separate between ordinary system traffic and attack traffic. This is the explanation behind utilizing the benchmark datasets for executing different attack discovery framework of this paper. The available attack datasets [14]-[17] are DARPA 1998, KDD Cup99, NSL KDD, UNSW NB15, etc. The DARPA 1998, KDD Cup99, and NSL KDD consists of 42 attributes including the class label. The UNSW NB15 dataset consists of 48 attributes including the class label.

C. Review Regarding Detection

Multiple detection methods have been carried out in various literatures. It includes traditional detection, ML-based and DL Neural-network based detection. In few research hybrid method is also used. Various detection techniques are analyzed in the following discussion.

D. Traditional Detection

A sandbox, in computer security, is a security component wherein a different, confined condition is made and in which several functions are restricted [18]. A sandbox is regularly utilized when untested code or entrusted programs from outsider sources are being utilized. Sandbox also has few constrain. Some sandbox apparatuses just deal with explicit sorts of PDF assaults like MD Scan for Java Script, [19] Nozzle for heap spraying [20], or it only records dynamic behavior of a system and still requires manual analysis to detect as in the case of CW Sandbox [21].

Huaibin Wang, HaiyunZhou, Chundong Wanghas discussed about VM-based different IDSs [22]. They have recommended to deploy VM-based numerous IDSs in each layer to observe specific virtual component. Additionally, they have also proposed the cloud alliance view, by the communication agents exchanging shared cautions commonly to withstand Denial of Service (DoS) and Distributed Denial-of-Service (DDoS). On this premise, they have accomplished an identity authentication of the communication agents, to improve the unwavering quality of the alarms. Through the evaluation of simulation results, the proposed device framework had a benefit for observing VMs on the detection rate.

E. Machine Learning (ML) and Deep Learning (DL) Based Detection

Machine learning algorithm learns from data [23]. Tom Mitchell precisely defines it as a computer program which learns from experience in respect to task and final outcome is the performance [24].

Vipin Kumar [25] used k mean clustering approach on NSLKDD dataset to perceive the accuracy for intrusion detection. Shilpaet Al [26] used fundamental element evaluation on NSLKDD dataset for feature selection and dimension pruning approach for evaluation on anomaly detection. In general, network intrusion detection has been broadly improved by applying data mining and machine learning technique, which has largely utilized individual conduct patterns from the community site visitors' data.

Support Vector Machine (SVM) is used, as a method in a study, to evaluate IDS [27]. Among various approaches of IDS, SVM acts as a classifier with false alarm and detection rate as a measure of performance. Authors in a study [28] used Markov Chain implementation as classifier and Apriori algorithm to remove isolated data from the database and also used to judge the performance of NIDS. K-Means, an unsupervised algorithm, is used for classification, defines an unlabeled class to which the clustering is performed.

III. Preliminaries

This section provides a brief background about network intrusion, and the four intelligent algorithm used in this study.

A. Concept of Network Intrusion

Modern technology has broken the border of digital intrusion and also digital threat. Attack in Estonia, Iran's nuclear power plant, digital espionage, financial damage-all of these are the newest threat of modern internet technology. Digital intrusion is the first step and the most common type of attack or threat [29]. Then onward malwares are injected or further important arsenals are used. Thereby, if intrusions are monitored and checked then first line defense can possibly be achieved.

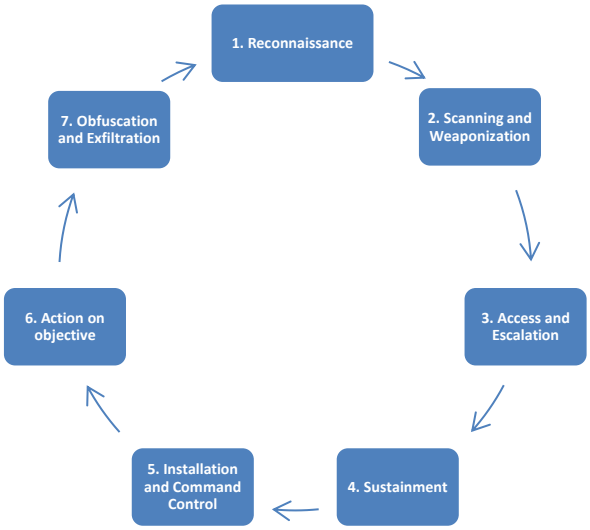

Fig. 1: Network Attack Cycle.

B. Concept of Network Attack

Analyzing the Fig. 1 it is clearly understood that first three steps of network attack cycle are related to intrusion. Therefore, it plays a vital role in overall attack cycle.

C. Overview Regarding Algorithm

There are many types of algorithms practiced in machine learning. But four suitable methods are utilized for performance analysis.

D. Classification and Regression Tree

Leo Bremen introduced the term CART. CART refers Decision Tree algorithm. It is used for classification or regression predictive modeling problems. Classically, this algorithm is mentioned as "decision trees". However, they are also mentioned by the more modern term CART, on some platforms like R [30].

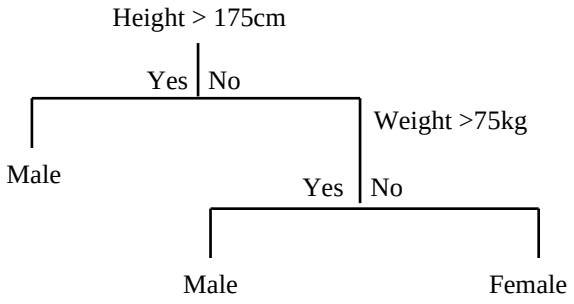

Fig. 2: Classification and Regression Tree.

The tree may be preserved to file as a graph or a set of rules. For example, it is given below as a set of rules.

- If Height > 175 cm, Then Male.
- If Height <= 175 cm AND Weight >75 kg, Then Male.
- If Height <= 175 cm AND Weight <= 75 kg, Then Female.
- Make Predictions with CART Models.

E. Random Forest

Random forest consists of a large number of individual decision trees. It operates as an ensemble. Each individual tree splits out a class prediction. The class with the most votes become model's prediction (see figure below) [31].

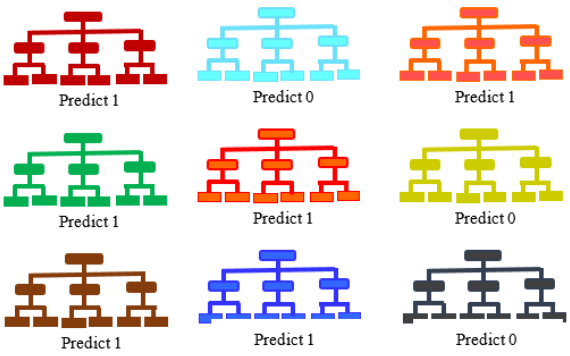

Fig. 3. Random Forest.

F. Naive Bayes Classifier

In the arena of Machine Learning, Naïve Bayes performs classification. The core essence of the classifier depends on Bayes theorem.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

This theorem follows probability of happening and occurs where the later one is evidence and the initial one is hypothesis. Predictors or features are independent which means one particular feature does not affect the other. Therefore, this technique is called naïve [32].

G. Multi-Layer Perception

Multi-layer perceptron is a deep learning technique where more than one linear layer (combination of **neurons**) is involved. In a three-layered network, first layer will be the *input layer* and last one will be *output layer* and with a hidden layer in between [33].

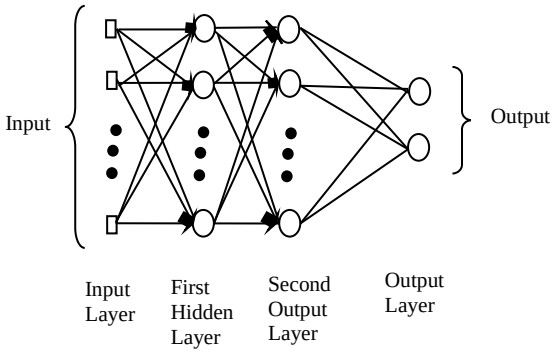

Fig. 4. *Multi-Layer Perception Model (MLP).*

IV. Experimental Results and Analysis

This section discusses the performance of the proposed IDS model which is furthermore analyzed with two experiments. In case of MLP, all the features are supplied as input to the neural network and then are trained using back-propagation. Then the accuracy is calculated from the test data. But in case of Random Forest, CART, and Naive Bayes, necessary parameters are utilized. However, in the second experiment, only the most important features are extracted, and then supplied as input along with the training dataset. Finally, the results obtained are presented in tabular as well as graphical form.

A. Application of ML Methods with Generalized Features

In the first experiment four methods are applied to find out accuracy in both normal flow of data and also for intrusion. Among all the ML Method, CART and MLP has provided a better accuracy. Whereas Random Forest has provided a greater intrusion detection which is 98.547%. In Fig. 5 graphical presentation has been displayed the same.

TABLE I: Test Accuracy for Normal Flow of Data and Intrusion Detection Using 41 Features

Type of Algorithm	Type of Data	Accuracy
Random Forest	Normal	85.387
	Intrusion	98.547
CART	Normal	99.086
	Intrusion	96.51
Naive Bayes	Normal	85.606
	Intrusion	93.265
MLP	Normal	93.387
	Intrusion	94.312

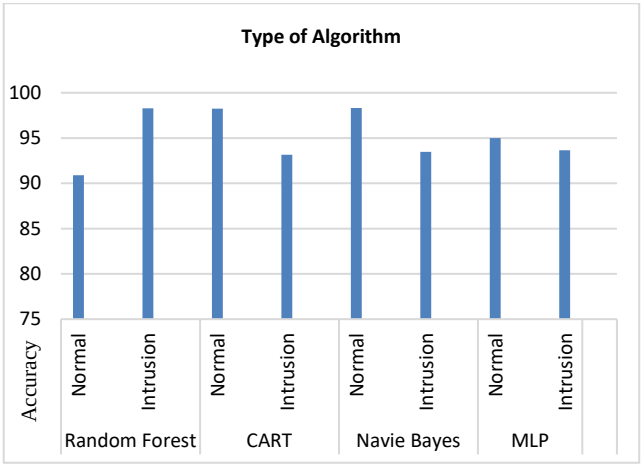

Fig. 5. Comparison among the ML and DL Methods for finding out Accuracy using Generic Features.

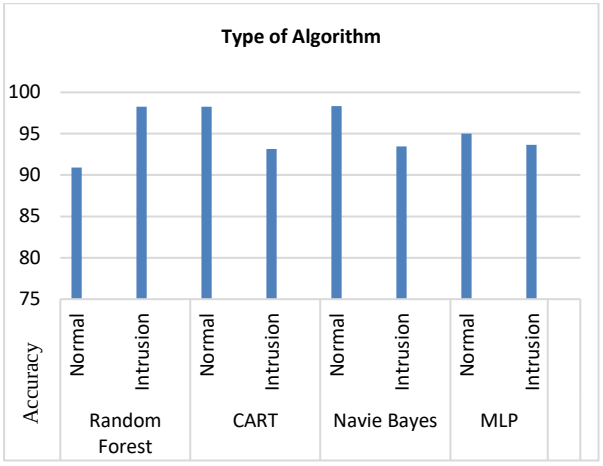

Fig. 6. Comparison among the ML and DL Methods for finding out Accuracy using Selective Features

B. Application of ML and DL Methods with Selective Features

Repeatedly in the second experiment, four methods are applied to find out accuracy in both normal flow of data and also for intrusion. But in this case selective features are applied in four different ML Methods like Random Forest, CART, Naive Bayes, & MLP. Here, CART and Naive Bayes has provided a better accuracy although Random Forest has also provided a better intrusion detection like previous which is 98.266%. Performance of MLP has also been displayed a significant improvement. In the Fig. 6 graphical presentation has also displayed the overall performance.

TABLE II: TEST ACCURACY FOR NORMAL FLOW OF DATA AND INTRUSION DETECTION USING SELECTIVE 15 FEATURES

Type of Algorithm	Type of Data	Accuracy
Random Forest	Normal	90.903
	Intrusion	98.266
CART	Normal	98.246
	Intrusion	93.167
Naive Bayes	Normal	98.331
	Intrusion	93.458
MLP	Normal	95.007
	Intrusion	93.652

C. Analytical Review

Experimental results in both cases have displayed reasonably good performance. Use of selective features and elimination of few less important parameters have also improved the overall performance. After analyzing overall results, Classification and Regression Tree is found to be a stable and better method keeping in mind that Random Forest provided the best intrusion detection in both cases.

TABLE III: COMPARISON OF ACCURACY AMONG ML AND DL METHODS USING GENERIC AND SELECTIVE FEATURES

Type of Algorithm	Type of Data	Accuracy Generic Features	Accuracy Selective Features
Random Forest	Normal	85.387	90.903
	Intrusion	98.547	98.266
CART	Normal	99.086	98.246
	Intrusion	96.51	93.167
Naive Bayes	Normal	85.606	98.331
	Intrusion	93.265	93.458
MLP	Normal	93.387	95.007
	Intrusion	94.312	93.652

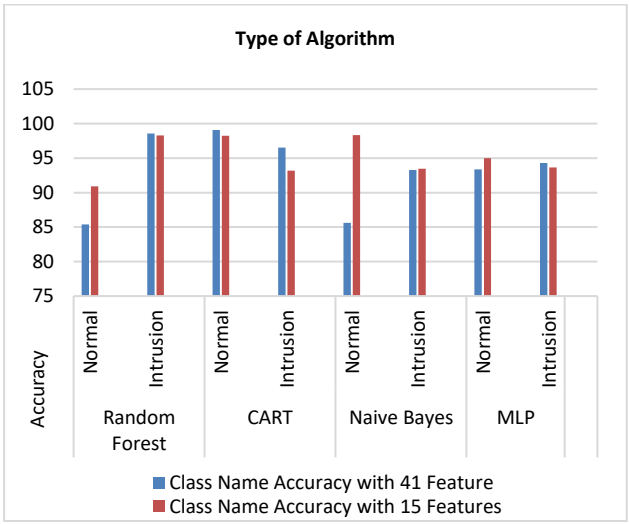

Fig. 7. Graphical Representation and Comparison of Accuracy among ML Methods using Generic and Selective Features.

D. Proposed IDS Model

A model consisting ML and DL method is proposed in Fig. 8. Here MLP with Back Propagation algorithm is used and Random Forest is taken as ML method. The selected methods are used considering the performance in accuracy.

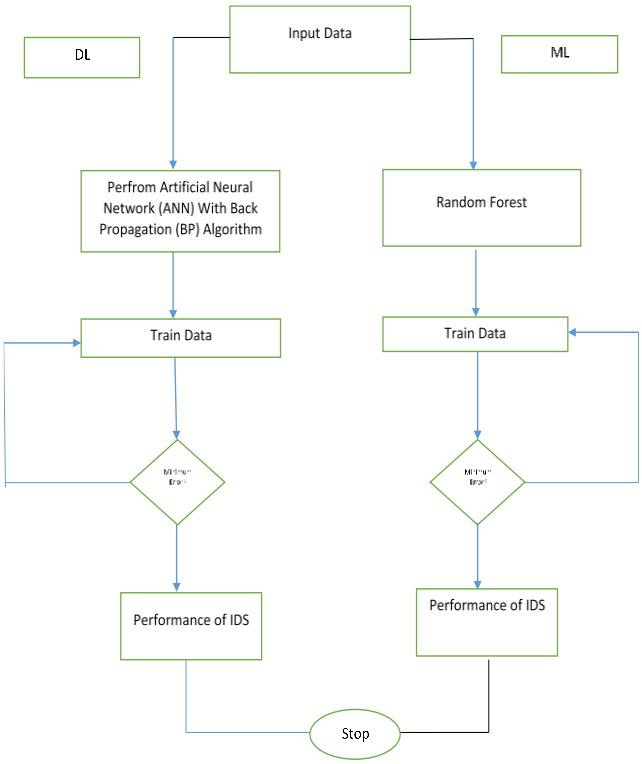

Fig. 8. Proposed IDS Model.

V. CONCLUSION

Sovereignty of a country is assured by ensuring the border security. But at the age of twenty first century border security is redundant where digital security is not guaranteed as digital world has got no border. However, living with modern technology is an important arsenal to ensure proper digital security. In this thesis, digital security vulnerabilities are discussed at the first place and subsequently it is manifested that the digital security technologies like malware removal programs, antivirus programs and firewalls, lack success to provide absolute protection. Attackers always employ updated techniques to assault the network, as well as its users. Wherefore, latest DL and ML methods are explored and finally necessary training and tests are carried out to measure the accuracy of the various DL and ML methods. MLP, one of the deep learning method along with various ML methods like Random Forest, CART, Naïve Bayes is also analyzed. The ML and DL methods are found to be very prudent in network security. Eventually, an IDS model is proposed where MLP and Random Forest is put into action.

REFERENCES

[1] C. E. Land Wehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws", ACM Comput. Surv. vol. 26, no .3, pp. 211–254, 1994.
[2] Olanrewaju R F, Khan B U I, Anwar F, Khan AR, Shaikh FA, Mir MS. "MANET– A cogitation of its design and security issues, "Middle-East Journal of Scientific Research. 2016;24(10):3094−107.
[3] KhamphakdeeN, BenjamasN, SaiyodS. "Network traffic data to ARFF converter for association rules technique of data mining," IEEE Conference on Open Systems (ICOS), IEEE;2014Oct.p.89−93.Crossref.
[4] Yu S C, Guo H, Yu G X, Jin X L, Zhang L N, Shao T J. "The solution to how to select an optimal set of features from many features used to intrusion detection system in wireless sensor network." 2010 Second WRI Global Congress on Intelligent Systems (GCIS), IEEE; 2010 Dec.3.p.368−71.Crossref.
[5] Olanrewaju R F, Habaebi M H. "Malicious behavior of node and its significant security techniques in MANET- A review," Australian Journal of Basic and Applied Sciences.2013;7(12):286−93.
[6] The history of intrusion detection systems (IDS) Part 1 Threat stack. Date accessed: 09/09/2015.https://www.threatstack.com/blog/the-history-of-intrusion-detection-systems-ids-part-1/.
[7] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier, "Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172–179, 2003.
[8] Anderson, J. A. (1995). "An introduction to Neural Networks," MIT Press.
[9] Rhodes, B. C., Mahaffey, J. A., &Cannady, J. D. (2000). "Multiple self-organizing maps for intrusion detection." In Proceedings of the 23rdnational information systems security conference (pp. 16-19).
[10] Al-Yaseen, W. L., Othman, Z. A., & Nazri, M. Z. A. (2017). "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system." Expert Systems with Applications, 67, 296-303.
[11] Chen, C. M., Chen, Y. L., & Lin, H. C. (2010). "An efficient network intrusion detection", Computer Communications, 33(4), 477-484.
[12] Deepa, A. J., &Kavitha, V. (2012). "A comprehensive survey on approaches to intrusion detection system." Procedia Engineering, 38,2063-2069.
[13] Thaseen, S., & Kumar, C. A. (2013). "An analysis of supervised tree based classifiers for intrusion detection system." In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on (pp. 294-299). IEEE.
[14] F. Iglesias, T. Zseby, "Analysis of network traffic features for anomaly detection," Machine Learning 101 (1-3) (2015) 59–84. doi:10.1007/525 s10994-014-5473.
[15] N. Moustafa, J. Slay, "The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set," Information Security Journal: A Global Perspective 25 (1-3) (2016) 18–31. doi:10.1080/19393555.2015.1125974.
[16] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in: Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, IEEE,2009, pp. 1–6. doi:10.1109/CISDA.2009.5356528.
[17] J. McHugh, "testing intrusion detection systems: a critique of the 1998 535 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory", ACM Transactions on Information and System Security(TISSEC) 3 (4) (2000) 262–294. doi:10.1145/382912.382923.
[18] www.techopedia Space issue.
[19] Z. Tzermias, G. Sykiotakis, M. Polychronakis, and E. P. Markatos, "Combining Static and Dynamic Analysis for the Detection of Malicious Documents, in Proceeding of the fourth Workshop on European Workshop on System Security," (Salzburg, Austria),2011.
[20] P. Ratanaworabhan, B. Livshits, and B. Zorn, "NOZZLE: A Defense Against Heap spraying Code Injection Attacks, in SSYM'09 Proceeding soft the 18th conference on USENIX security symposium," (Berkeley, CAUSA), 2009.
[21] C. Willems, T. Holz, and F. Freiling, "Toward Automated Dynamic Malware Analysis Using CW Sandbox".
[22] Huaibin Wang, Haiyun Zhou, ChundongWang "Virtual Machine-based Intrusion Detection System Framework in Cloud Computing Environment" JCP 2012 Vol.7(10): 2397-2403 ISSN: 1796-203Xdoi: 10.4304/jcp.7.10.2397-2403.
[23] I. Good Fellow, Y. Bengio, and A. Courville, "Deep Learning," The MIT Press, 2016.
[24] T. Mitchell, "Machine Learning," McGrawHill,1997.
[25] Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset" International Journal of Soft Computing and Engineering (IJSCE)ISSN:2231-2307, Volume-3, Issue-4, September2013.
[26] Shilpalakhina, Sini Joseph and Bhupendraverma, "Feature Reductiousing Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol.2(6),2010,1790-1799.
[27] Mohammadpour L, Hussain M, Aryanfar A, Raee VM, Sattar F. "Evaluating performance of intrusion detection system using support vector machines," International Journal of Security and Its Applications. 2015 Sep; 9 (9): 225−34. Cross ref.
[28] Brindasri S, Saravanan K. "Evaluation of network intrusion detection using Markov chain, "International Journal on Cybernetics and Informatics (IJCI).2014Apr; 3 (2): 11−20. Crossref.

[29] Clarence Chio and David Freeman, "Machine Learning and Security," O'REILLY, P.6.
[30] https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/ Accessed on 25 Aug 2020.
[31] https://towardsdatascience.com/ understanding-random-forest-58381e0602d2Accessed on 25 Aug 2020.
[32] https://towardsdatascience.com/ naive-bayes-classifier-81d512f50a7c.
[33] https://medium.com/@xzz201920/multi-layer-perceptron-mlp-4e5c020fd28aAccessed on 25 Aug 2020.

Shah Md. Istiaque is a military communication expert. As part of his career requirement, he completed graduation from the Department of Electrical Electronics and Communication Engineering from Military Institute of Science and Technology, Dhaka. Presently, he is undergoing a research programme on "Smart Intrusion Detection System Comprised of Machine Learning and Deep Learning" as part of Masters Programme in Information Security System, in Bangladesh University of Professionals (BUP). He is also one of the authors of an international publication, published in IEEE Digital Explorer titled "Design and integrate dual renewable energy in a residential building of urban area: A step towards the self-sustained smart energy system for Bangladesh"DOI: 10.1109/ICEEICT.2014.6919139.

Asif Iqbal Khan earned his Bachelor of Science in Engineering degree Information and Communication Technology from Mawlana Bhashani Science and Technology University in 2019. His major was Information and Communication Technology. Currently he is pursuing his Master's degree in the same institution. His research interests are in the field of Artificial Intelligence, Machine Learning, Data Science, Natural Language Processing, Predictive analysis etc. He specializes in data analysis with tools like python, R etc.

Sajjad Waheed earned his doctorate degree on Computer Engineering from the Istanbul University in 2013. He is currently working as a Professor in the Department of Information and Communication Technology in the Mawlana Bhashani Science and Technology University, Bangladesh. His research interests are in the field of artificial systems, machine learning, cryptography, data analysis, system development, etc. He has received funds for his research projects from the government and the university. He has published more than twenty research papers in internationally acclaimed peer-reviewed journals related to his research works.